Peer Assessment Using Aropä

John Hamer **Catherine Kell** **Fiona Spence**

Department of Computer Science   Centre for Flexible and Distance Learning

The University of Auckland

J.Hamer@cs.auckland.ac.nz c.kell@auckland.ac.nz f.spence@auckland.ac.nz

Abstract

Aropä is a web-based peer assessment support tool that has been used extensively in a wide variety of settings over the past three years. We describe the design of Aropä and how it can be configured, and present some results from a research study into the use of peer assessment in large undergraduate courses. There is evidence to show that while students find peer assessment challenging, it can be an effective aid to learning. The study also reveals marked differences in attitude toward peer assessment between different student bodies.

1 Introduction

Over the past three years we have introduced peer assessment into a variety of undergraduate courses in Computer Science, Software Engineering, Pharmacology, English and Photography. The courses range in size from forty to four hundred, and in level from introductory to fourth year. The material assessed includes computer programs, technical reports, software designs, academic and mixed-mode essays, photographs, posters, team member performances, and presentations. Most material and most reviews were prepared individually, but we also have instances of team assignments and team reviewing. The assessment exercises were primarily formative, with most including a token mark for participation.

It was only possible to conduct peer assessment exercises on this scale with the help of support software, in this case a locally-developed web-based tool called Aropä (*Aropä* means "peer review" in Maōri). A detailed description of Aropä has not previously appeared in the literature, so we take the opportunity in section 3 to present an overview of the tool and its appearance from a student perspective. We then present observations on three of the courses that have used the tool. The observations were taken from anonymous surveys and interviews with both teaching staff and students. The results show that peer assessment can aid learning in a variety of ways. We also observe that marked differences are evident between different student groups.

2 Why peer assess?

Peer assessment is attracting increasing attention from educators looking for new ways of improving learning outcomes in undergraduate courses. Many of the tasks associated with peer assessment are associated with Bloom's 1956 "higher" learning outcomes of *analysis* and *evaluation*. More specifically, literature surveys by Ballantyne et al. (2002) and Topping (1998) suggests that peer assessment can:

- help to consolidate, reinforce and deepen understanding, by engaging students in cognitively demanding tasks: reviewing, summarising, clarifying, giving feedback, diagnosing misconceptions, identifying missing knowledge, and considering deviations from the ideal;
- highlight the importance of presenting work in a clear and logical fashion;
- expose students to a variety of styles, techniques, ideas and abilities, in a spectrum of quality from mistakes to exemplars;
- provide feedback swiftly and in quantity. Feedback is associated with more effective learning in a variety of settings. Even if the quality of feedback is lower than from professional staff, its immediacy, frequency and volume may compensate;
- promote social and professional skills;
- improve understanding and self-confidence; and
- encourage reflection on course objectives and the purpose of the assessment task.

Historically, peer assessment has been largely confined to small graduate courses or in tutoring contexts. However, the potential benefits for large undergraduate classes are considerable. In addition to the suggested learning benefits, time saving is also often given as a pragmatic reason in favour of peer assessment (Ballantyne et al. 2002).

3 Aropä

The traditional form of course assignment work involves students working independently to prepare a submission, which is then marked by a course marker who produces a grade and (perhaps) some feedback for the student to reflect on (figure 1).

Peer assessment modifies this cycle to involve the student in reviewing and possibly rating the feedback (figure 2). New "dispute" and "rating" loops are introduced by peer assessment, in acknowledgement that the reviewing will sometimes be flawed. While reviewing errors can and do arise in both kinds of activity, peer assessment involves a change in power relations between author and reviewer that legitimises questioning a review.

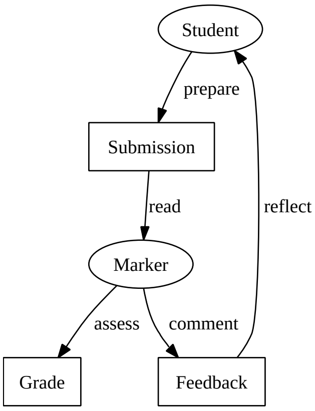


Figure 1: The traditional course assignment cycle

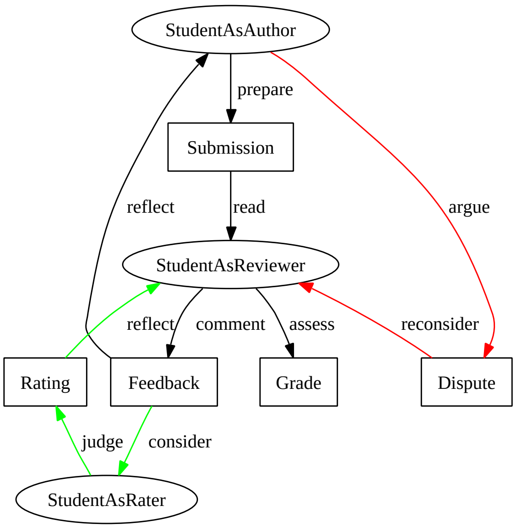
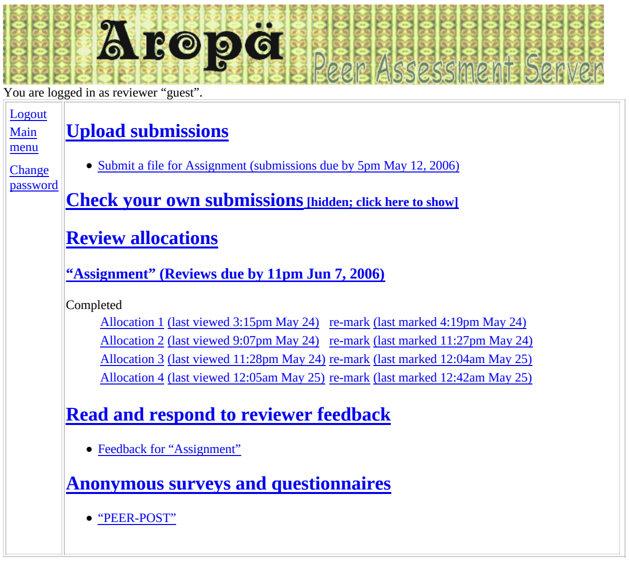


Figure 2: The peer-assessment cycle

The main functions of Aropä are to manage: allocation of submissions to reviewers; web-entry of reviews; access to feedback, disputes and ratings; and weighted-average grade calculation.

3.1 The student interface

A student assumes multiple roles during peer assessment, first as a producer of material to be assessed, then as a reviewer, and finally as a receiver of feedback. Aropä presents all these roles to the student simultaneously, arranged in sections that (in a Javascript-enabled browser) can be shown or hidden by clicking on the section title.



You are logged in as reviewer “guest”.

Logout
Main menu
Change password

Upload submissions

- Submit a file for Assignment (submissions due by 5pm May 12, 2006)

Check your own submissions [hidden; click here to show]

Review allocations

“Assignment” (Reviews due by 11pm Jun 7, 2006)

Completed

Allocation 1 (last viewed 3:15pm May 24) re-mark (last marked 4:19pm May 24)
Allocation 2 (last viewed 9:07pm May 24) re-mark (last marked 11:27pm May 24)
Allocation 3 (last viewed 11:28pm May 24) re-mark (last marked 12:04am May 25)
Allocation 4 (last viewed 12:05am May 25) re-mark (last marked 12:42am May 25)

Read and respond to reviewer feedback

- Feedback for “Assignment”

Anonymous surveys and questionnaires

- “PEER-POST”

Figure 3: The student interface to Aropä

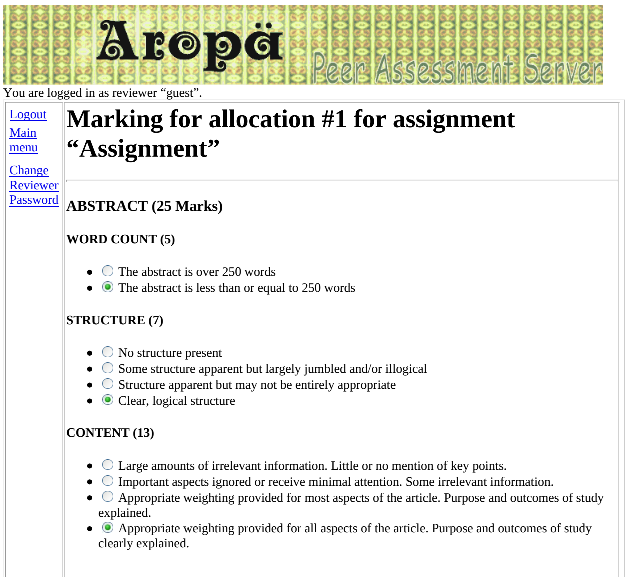
Figure 3 shows a typical main screen after log in. The left of the screen displays links to logout, change password, and (redundantly) return to the main menu. On the right are sections to upload files for any current assignments, check that files have been uploaded correctly, review allocated assignments, read feedback, and participate in any class surveys. It is possible for a student to have more than one peer assessed assignment at a given time, in which case each assignment appears in a subsection, and an additional section is displayed at the top of the screen to allow a subset of the current assignments to be displayed.

Uploading is a straightforward operation. The student is prompted for the name of one or more files, which are sent to the web server when a “save” link is invoked. Currently no constraints are imposed on the names, types, size or number of files. Some courses (particularly in Computer Science) have their own submission system, in which case Aropä can be instructed to look for submitted files in a directory mounted on the server.

The “Check your own submissions” section is provided mainly to reassure students that their files have been received by Aropä. One of the biggest technical problems we have faced is when students submit files (from home) in a non-standard document format (typically old versions of MacWrite or Lotus Notes, etc.). This facility allows students to confirm their submissions are readable in a “standard” environment, such as a University computer laboratory.

Students spend the largest portion of their time in the “Review allocations” section. This section contains links to view and mark each of the allocations they have been assigned to review. Usually reviewing is double-blind, in which case Aropä numbers the allocations sequentially. If half-blind reviews are used, the name of the author is displayed instead. The "view" link displays a list of all the files submitted by the author. The files can be downloaded individually, or together as a compressed "zip" archive.

Marking is done using a "grading rubric" prepared by the lecturer. The rubric can contain HTML elements (headings, paragraphs, lists, tables, etc.) together with checkboxes, groups of radio buttons, and text areas. Figure 4 shows part of a rubric used for grading a report on a pharmaceutical drug. Three groups of radio buttons are shown. This rubric included eleven such groups in total, followed by comment areas for "what did you like best about the assignment," "how could it be improved," and "any general comments."



You are logged in as reviewer "guest".

Marking for allocation #1 for assignment "Assignment"

ABSTRACT (25 Marks)

WORD COUNT (5)

- ○ The abstract is over 250 words
- ● The abstract is less than or equal to 250 words

STRUCTURE (7)

- ○ No structure present
- ○ Some structure apparent but largely jumbled and/or illogical
- ○ Structure apparent but may not be entirely appropriate
- ● Clear, logical structure

CONTENT (13)

- ○ Large amounts of irrelevant information. Little or no mention of key points.
- ○ Important aspects ignored or receive minimal attention. Some irrelevant information.
- ○ Appropriate weighting provided for most aspects of the article. Purpose and outcomes of study explained.
- ● Appropriate weighting provided for all aspects of the article. Purpose and outcomes of study clearly explained.

Figure 4: A formative grading rubric used in Pharmacology

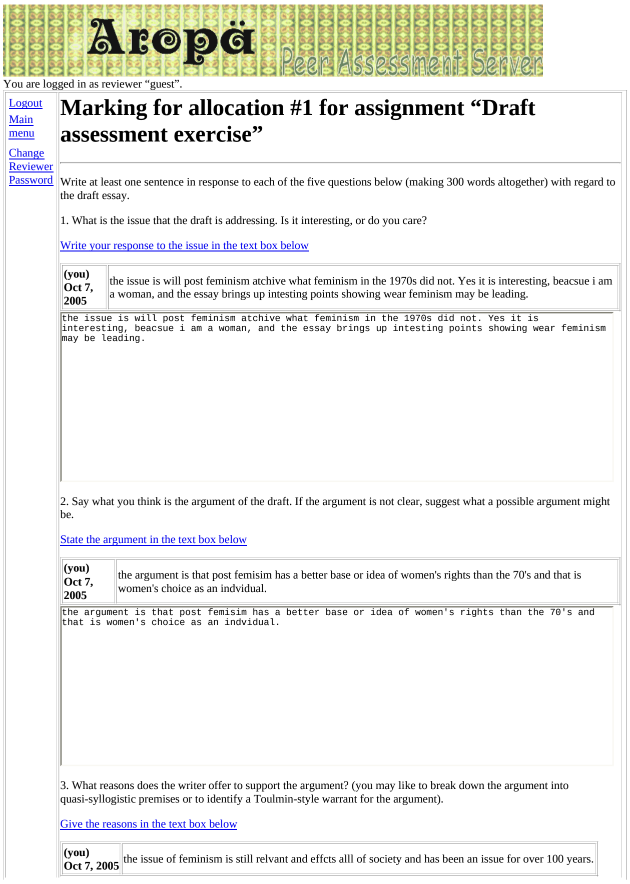
A contrasting rubric used in an English class is (partly) shown in figure 5. This rubric is entirely formative, consisting of a series of text boxes but no "grading" elements.

In general, rubrics fall between these two extremes, with grading elements intermingled with free-form comment areas.

Facilities are provided for formatting comments. This can be done with a simple "wiki markup" or (on a modern browser) using a Javascript HTML editor[1]. The Javascript editor was added in the first 2006 semester, and appears to have resulted in a significant increase in the amount of written feedback. We return to this issue in section 4.2.

Feedback is provided to the student through a modified version of the grading rubric. The radio and check buttons are replaced by the number of reviewers awarding each mark. The screen shot in figure 6 shows that all three reviewers agreed that the notes were comprehensive, and two out of three approved of the organisation and conventions. Comments are shown next to a reviewer identifier (see figure 7, which lets the student reconcile comments made in different comment boxes.

The final facility provided by Aropä is an on-line survey. Teaching staff can set up one or more surveys soliciting anonymous feedback from students. The surveys use the same elements as rubrics. Surveys


Aropä
Peer Assessment Server
Logout
Main menu
Change Reviewer Password
Write your response to the issue in the text box below
State the argument in the text box below
Give the reasons in the text box below

You are logged in as reviewer "guest".

Marking for allocation #1 for assignment "Draft assessment exercise"

Write at least one sentence in response to each of the five questions below (making 300 words altogether) with regard to the draft essay.

1. What is the issue that the draft is addressing. Is it interesting, or do you care?

(you) **Oct 7, 2005**	the issue is will post feminism atchive what feminism in the 1970s did not. Yes it is interesting, beacsue i am a woman, and the essay brings up intesting points showing wear feminism may be leading.

```
the issue is will post feminism atchive what feminism in the 1970s did not. Yes it is
interesting, beacsue i am a woman, and the essay brings up intesting points showing wear feminism
may be leading.
```

2. Say what you think is the argument of the draft. If the argument is not clear, suggest what a possible argument might be.

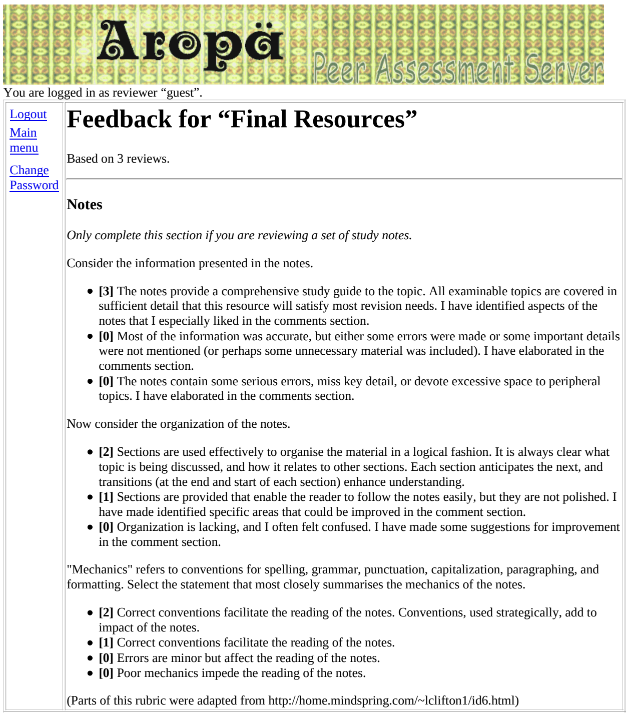
(you) **Oct 7, 2005**	the argument is that post femisim has a better base or idea of women's rights than the 70's and that is women's choice as an indvidual.

```
the argument is that post femisim has a better base or idea of women's rights than the 70's and
that is women's choice as an indvidual.
```

3. What reasons does the writer offer to support the argument? (you may like to break down the argument into quasi-syllogistic premises or to identify a Toulmin-style warrant for the argument).

(you) **Oct 7, 2005**	the issue of feminism is still relvant and effcts alll of society and has been an issue for over 100 years.

Figure 5: A formative grading rubric used in English



You are logged in as reviewer "guest".

Feedback for "Final Resources"

Based on 3 reviews.

Notes

Only complete this section if you are reviewing a set of study notes.

Consider the information presented in the notes.

- **[3]** The notes provide a comprehensive study guide to the topic. All examinable topics are covered in sufficient detail that this resource will satisfy most revision needs. I have identified aspects of the notes that I especially liked in the comments section.
- **[0]** Most of the information was accurate, but either some errors were made or some important details were not mentioned (or perhaps some unnecessary material was included). I have elaborated in the comments section.
- **[0]** The notes contain some serious errors, miss key detail, or devote excessive space to peripheral topics. I have elaborated in the comments section.

Now consider the organization of the notes.

- **[2]** Sections are used effectively to organise the material in a logical fashion. It is always clear what topic is being discussed, and how it relates to other sections. Each section anticipates the next, and transitions (at the end and start of each section) enhance understanding.
- **[1]** Sections are provided that enable the reader to follow the notes easily, but they are not polished. I have made identified specific areas that could be improved in the comment section.
- **[0]** Organization is lacking, and I often felt confused. I have made some suggestions for improvement in the comment section.

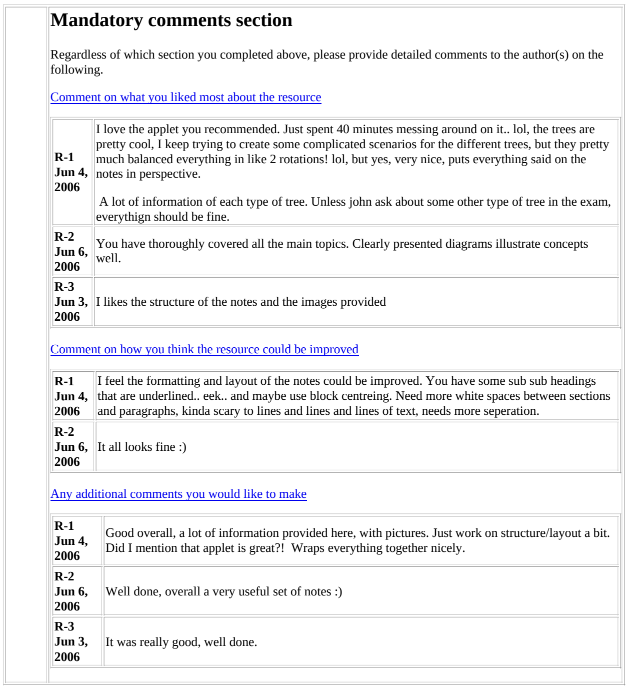
"Mechanics" refers to conventions for spelling, grammar, punctuation, capitalization, paragraphing, and formatting. Select the statement that most closely summarises the mechanics of the notes.

- **[2]** Correct conventions facilitate the reading of the notes. Conventions, used strategically, add to impact of the notes.
- **[1]** Correct conventions facilitate the reading of the notes.
- **[0]** Errors are minor but affect the reading of the notes.
- **[0]** Poor mechanics impede the reading of the notes.

(Parts of this rubric were adapted from http://home.mindspring.com/~lclifton1/id6.html)

Figure 6: Reviewer grade feedback

[1] We use the tinymce editor (Moxiecode Systems AB 2006)



Mandatory comments section

Regardless of which section you completed above, please provide detailed comments to the author(s) on the following.

Comment on what you liked most about the resource

R-1 **Jun 4, 2006**	I love the applet you recommended. Just spent 40 minutes messing around on it.. lol, the trees are pretty cool, I keep trying to create some complicated scenarios for the different trees, but they pretty much balanced everything in like 2 rotations! lol, but yes, very nice, puts everything said on the notes in perspective. A lot of information of each type of tree. Unless john ask about some other type of tree in the exam, everythign should be fine.
R-2 **Jun 6, 2006**	You have thoroughly covered all the main topics. Clearly presented diagrams illustrate concepts well.
R-3 **Jun 3, 2006**	I likes the structure of the notes and the images provided

Comment on how you think the resource could be improved

R-1 **Jun 4, 2006**	I feel the formatting and layout of the notes could be improved. You have some sub sub headings that are underlined.. eek.. and maybe use block centreing. Need more white spaces between sections and paragraphs, kinda scary to lines and lines and lines of text, needs more seperation.
R-2 **Jun 6, 2006**	It all looks fine :)

Any additional comments you would like to make

R-1 **Jun 4, 2006**	Good overall, a lot of information provided here, with pictures. Just work on structure/layout a bit. Did I mention that applet is great?! Wraps everything together nicely.
R-2 **Jun 6, 2006**	Well done, overall a very useful set of notes :)
R-3 **Jun 3, 2006**	It was really good, well done.

Figure 7: Review comment feedback

can be open to all students, or limited to participants from one or more assignments or any ad-hoc group.

3.2 Configuration

A typology of peer assessment from (Topping 1998) is reproduced in table 1. Aropä has been designed to support all of the variations in this typology, and we have experience in applying the system to many different configurations.

We have already made mention of the range of Subject areas that have used Aropä, and will give examples of different staff and student Objectives, Staff assessment, Official weight Place, Time, Requirement and Reward in the section 4.

The rubric format described in section 3.1 supports a mixture of both quantitative and qualitative Focus.

The Product being assessed is most often written work of some form or other, but this is not a restriction. Aropä has also been used to assess oral presentations and for team member performance reviews. In such cases, photographs of the assessee are uploaded, to remind the reviewer of which presentation or team member they are reviewing for a given allocation.

The variables Directionality, Year and Ability all concern the allocation of reviews to reviewers. Aropä allocations can be manual or automatic (i.e., randomly generated). Automatic allocations can be made in one or more *streams*, where each stream involves a subset of either the reviewers or assessees. For example, a class can be grouped into top, middle and poorly performing students. Each submission can then be assigned to one or more reviewers from each group. Or, conversely, each reviewer can be allocated one or more submissions from each group. Streaming has been used in this way in some Computer Science courses.

Reviewers and assessees are independent sets, so there is no difficulty in allocating students from different courses or years. To date, this has been done with staff markers using the system to mark student work, and with a final-year Software Engineering class reviewing work from a third-year design course.

Self-review is supported as a per-activity option. It has been occasionally used in conjunction with peer review; we have no experience with using self-review in isolation. However, positive findings on self-review have been reported in Dochy et al. (1999), and we plan to investigate this further in the future.

Submissions can also be "seeded" with pre-prepared solutions. These may be model solutions, or solutions with particular flaws. Seeded submissions are generally not identified to the reviewers. While it would be possible to use seeded submissions as a quality check, this has not been done in any of the courses using Aropä. Rather, the motivation has been to ensure each student is exposed to at least one high quality solution.

Aropä supports individual or group submissions and reviews, in any combination. Groups used for both submission and review are not required have the same membership, although this is normally the case.

3.3 Other features

Review of reviews Aropä regards each peer assessment activity as an assignment in itself, one in which the reviewers are authors of their reviews. Reviews can thus be reviewed using all the facilities available to normal assignments[2]. Typically, "review reviews" are carried out by a lecturer or course tutor. The reviewer "quality" grades from a review-of-the-reviews are used to automatically weight the grades assigned by the reviewer, so the marks from "good" reviewers can be made to count for more.

Late submissions Reviewer allocations are most often made after all submissions have been received, so that only students who submit material for the activity are included as reviewers. Aropä keeps track of the time each allocation is viewed, and can switch allocations that have not been seen to make room for occasional late submissions.

Dialogue Double-blind dialogue between author and reviewer is supported. Author can respond to reviewer's comments during a "feedback" period. The administrator can decide whether this period overlaps with the review period (in which case a reviewer has the opportunity to change her comments and/or grades) or to keep the periods distinct.

No class lists Aropä does not connect to an enrolment database, and has no notion of a class list. For each activity, the participating students are determined either from the names on the submitted files or from a list prepared by the lecturer. This design makes it possible to include students who wish to participate in the activity, but are for some reason not enrolled, or whom the lecturer wishes to hide (such as the author of a seeded solution).

Grade calculation The grade for a submission is calculated using a weighted average of all the reviews. The reviewer weights can be assigned by the "review review" or using an automatic calibration algorithm (Hamer et al. 2005). A grade variance report identifies areas of significant disagreement between reviewers. The administrator can investigate, and exclude particular reviews or reviewers, or even rubric questions.

[2] The process can continue indefinitely. The grades for the final unreviewed reviews are computed using the algorithms described in Hamer et al. (2005)

Variable	Range of variation
Curriculum area/subject	All
Objectives	Of staff and/or students? Time saving or cognitive/affective gains?
Focus	Quantitative/summative or qualitative/formative or both?
Product/output	Tests/marks/grades or writing or oral presentations or other skilled behaviours?
Staff assessment	Substitutional or supplementary?
Official weight	Contributing to assessee final official grade or not?
Directionality	One-way, reciprocal, mutual?
Privacy	Anonymous/confidential/public?
Contact	Distance or face-to-face?
Year	Same or cross year of study?
Ability	Same or cross ability?
Constellation Assessors	Individual or pairs or groups?
Constellation Assessed	Individual or pairs or groups?
Place	In/out of class?
Time	Class time/free time/informally?
Requirement	Compulsory or voluntary for assessors/ees?
Reward	Course credit or other incentives or reinforcement for participartion?

Table 1: A typology of peer assessment, from (Topping 1998)

Persona An course administrator can temporarily assume the persona of a student at any time, without requiring to know their password. This feature has proved suprisingly useful for investigating reported problems.

Time-on-task A report calculates the approximate times when a student is using Aropä and for how long. The data is approximate because it cannot measure time spent away from the computer, but it does give an indication of the workload imposed by the peer assessment activity.

4 Courses using Aropä

4.1 Pharmacology

A large year-2 class in Pharmacology used Aropä in the second semester of 2005 and again (as a year-3 class) in the first semester of 2006. In 2005, 335 student participated, and in 2006 there were 180 essays submitted. In both instances the students worked individually writing an essay on a pharmaceutical drug, chosen from a list of six. The essays were around 4–8 pages, and followed a strict academic format (abstract, references, etc.). We report here on the 2005 exercise.

Students were instructed to hand in their assignments electronically to the Aropä website, and also to submit a second version (without reference list) to Turnitin (Barrie 2006), a plagiarism detection site. Assignments were then allocated randomly for students to mark, 4 per student, over a three day period. A detailed marking rubric was provided on the website (this was available for a week prior to the deadline), and marks were submitted via this rubric. Students were awarded 4% of their grade for marking the four assignments. The assignment itself contributed 20% of their final grade.

Students were asked to provide feedback on their understanding and prior perceptions about peer assessment. Only seven students responded to these initial questions. They all had some idea of what peer assessment was, although none had previously participated in peer assessment. All of these students expressed a reluctance to participate in this peer assessment process.

4.1.1 Positive feedback from the class representative and students

Positive feedback from the class representative[3] and the class indicated that the students had perceived and experienced two important pedagogical benefits of peer assessment.

Firstly, they had gained knowledge and understanding of the academic environment and processes. They felt they had gained a much better understanding of what is involved in marking generally and how to approach writing essays differently as a result of the experience: "*it was a good way to learn alternative ways of writing reports*"; "*I realised how poor some reports are*"; "*it was interesting to see how bad mine was compared to others.*"

Secondly, in relation to the actual content knowledge in the field of Pharmacology, students felt they had a deeper understanding of the drug topic, that it had reinforced learning and that they had been able to pick up on information about the drug which hadn't been found while researching for their own essay. There was a suggestion that it would be useful to be able to view reports on other drugs (not to mark) to further their understanding in Pharmacology generally.

In commenting positively on the marking process two students suggested that it would be good to be provided with three model answers (excellent, average and poor) so that you could benchmark your marking. There was some sense of surprise: "*that reviewers commented on the good points and the bad,*" and "*I didn't expect the marking to be done so fairly and consistently — students marked really well*"; "*I felt compelled to write comments... also to improve the students' work in the future.*"

There was little feedback on the interface except for a response saying that "*comment box at the side was a really good way to learn as lecturers don't usually have much time for comments.*"

[3] The University of Auckland operates a "class representative" system in which one or two students from each course volunteer to liaise with the department and raise any issues of concern.

4.1.2 Problems and recommendations from the class representative and students

Some comments related to details about the interface and technical issues. One asked "*How did Aropa generate a number out of four reviewers? How does it deal with bias?*" Students reported feeling anxious and as "*needing to feel safe about loading their stuff.*" They did experience some glitches, for example, "*when an assignment was submitted but it wasn't loaded*" and it would "*allay nerves if the system sent out a confirmation email after submission was made.*"

On pedagogical issues, the class representative's and students' feedback covered a number of areas, including:

- preparation
- the marking rubric
- 'genre' and language considerations
- the requirement for marking
- timing.

Each of these is detailed below.

The class representative and students reported that the exercise could have been improved with better preparation and more warning given to students to allay fears and anxieties. Students could have been better informed about the benefits of the approach ("*more theoretical justification of the use of peer assessment*") and how it had been successfully used in other contexts. Better participation in reviewing could possibly have been achieved through "*more encouragement.*"

Secondly, there were a number of suggestions about the marking process and the marking rubric. Although these indicate a sense of frustration on the part of students with this particular exercise, they can be seen as positive in the sense that they engage with what peer assessment is about and make positive suggestions to improve it. The marking rubric was seen by a number of students as too limiting, and needing to be expanded to include more than three options. This seemed to be indicating that students felt that the rubric favoured short answers tailored to the rubric, rather than the answers of students who had spent more time on researching and checking information. The class representative reported that students had felt that: "*The marking rubric did not account for the effort put into understanding the drug and its mechanisms — the rubric encouraged fitting information to it.*" It was also seen as not entirely correlating to the questions given previously — it did not correspond well to what was asked for in the assignment. There was a request for "*exceptionally well done*" criteria or "*bonus marks for deep level of research,*" and frustration that it only "*allowed you to give OK or good results.*" It was suggested that the marking rubric should cover spelling and grammar, and that marks should be allocated to "*essay format and quality.*"

Thirdly, although the assignment required a strict essay format with an abstract and referencing, there were interesting responses around this. Two students indicated that the class mixed BSc students with BPharm students and that there were "*different approaches, writing styles and opinions as to which sections were more important and deserved more focus*":

> *an essay by a BSc student and marked by a BPharm student could have been perceived very differently. This could have contributed to inconsistencies.*

Responses	Time spent
(6)	2 hrs ($\frac{1}{2}$ hr each)
(6)	6 hrs ($1\frac{1}{2}$ hrs each)
(5)	4 hrs total
(2)	1 hr total
(2)	50 minutes each assignment
(1)	45 mins each assignment
(1)	2 hrs per assignment
(1)	read the essays 3–4 times each (no time specified)
(1)	$1\frac{1}{2}$ days
(1)	not much

Table 2: Time spent marking assignments

The class representative noted that "*students need guidance as to how to deal with writing when English is clearly a second language.*"

More importantly, a number felt that students actually marking each other was unfair, and they felt uncomfortable with this contributing to their final marks. Two students indicate that they felt it wasn't always clear why marks were lost, one felt that "*everyone is competing so they would mark more harshly.*" One commented that: "*Real lecturers do not drag you down for minor things.*" An important comment was: "*It was difficult to tell whether what a person had written was correct if you hadn't found that information yourself.*"

Eighteen students said they were happy with the grades they received, while ten were dissatisfied. Three did not respond to that question.

Four responded that the exercise was a waste of time being scheduled just before tests and other assignments that needed attention. A number of students felt that they had not had enough time to do the marking. Nearly all the respondents noted that they took considerable time and made an effort to provide comments and feedback in their marking. But correspondingly they nearly all said they received very little or no feedback on their returned essays. One suggested that six hours was a lot of time and that they should perhaps have had three to mark rather than four. There was a range of reported time spent marking assignments (see Table 2) from "*not much*" to 1.5 days (five did not respond).

4.1.3 Mark reliability

Approximately twenty five assignments were randomly selected and cross checked by departmental staff. With one exception, all assignment marks varied by less than 12%; much of the discrepancy was due to differences in referencing interpretations. Twelve more assignments were remarked upon request of students. For the majority, the new mark differed by ±5% from the original mark, although there were three where the difference between marks was more than 10%. It was common for some or all markers to have given different marks for specific sections, but typically their overall marks agreed very closely.

4.2 English

A medium-size class in English used Aropä in the second semester of 2005 and again in the first semester of 2006. The course was called "Rhetoric in Public Culture." Students worked individually on a topic selected from six or seven broad subject areas. The peer assessment exercise was a formative review of a draft. A small number of marks were awarded for participating in the activity; no attempt was made to distinguish the quality of the participation. In both cases 80 students participated. Students were allocated three reviews each, and were given the option to write a self-review. Allocations were assigned randomly.

In writing their essays students were required to think carefully about the intended audience. The course lecturer was very interested in the power relations. He mentioned that "*the student has to feel sufficiently authoritative*" to comment and that this authority was created for them by the peer assessment task, in that they were the public audience. It seemed that many of the students who enrolled in this course aspired to careers as critics, reviewers, editors, opinion columnists, etc. Many of the students therefore had an interest in the genre of the review itself.

The lecturer expressed reservations about drawing up a rubric prior to the exercise:

> *as with filling in [academic performance review] forms, it organises you to mentally deal with the box. What happens if you find yourself in between boxes? If you are a creative person this is very likely to happen.*

Furthermore, he argued that:

> *while I am keen to make things transparent, the thing is organic, if you get a mechanical object it's an untrue reflection. Students learn about writing over the course of a whole degree.... A template can lead you to believe that there is a mechanical process involved, that writing a piece is an assemblage of things. What happens to criticality, creativity?*

In the end, he drew up the following rubric (part of this can be seen in the screen shot in Figure 5):

- Write at least one sentence in response to each of the five questions below (making 300 words altogether) with regard to the draft essay.
- What is the issue that the draft is addressing? Is it interesting, or do you care?
- Say what you think is the argument of the draft. If the argument is not clear, suggest what a possible argument might be.
- What reasons does the writer offer to support the argument? (You may like to break down the argument into quasi-syllogistic premises or to identify a Toulmin-style warrant for the argument).
- Suggest a counterargument to the argument of the draft. This comment may, alternatively, point out unexamined assumptions and/or missing or unacknowledged evidence.
- Identify a characteristic sentence of the writer. Say what you think is good about this sentence, or how this sentence can be improved (your chosen sentence may simply identify a repeated writing fault)

These questions show that one of the problems raised in the Pharmacology course had been addressed by this lecturer; i.e., the close relation between the requirements of the assignment and the organisation of the rubric.

4.2.1 Survey and interviews

All students were invited to fill in an on-line survey, both before and after the exercise. Twelve students completed the survey prior to the exercise, and ten answered after (four answered both before and after). We also interviewed eight students before and again after the 2005 exercise. These volunteers were self-selected, seven female and one male.

The fact that the rubric developed by the lecturer was integrally connected with the purpose of the assignment, combined with the fact that the students were not doing summative marking of each others' work, and that they were able to improve their essays by drawing on the peer reviews, led to quite different data from the previous example. In general it would be very difficult to divide up those responses which were negative and those which were positive. We will, however, attempt to address the same themes as in the previous example.

On the issue of preparation, responses indicated that the students were well-informed about the purpose of the exercise and felt relaxed about what their role was to be. The responses that we received seemed to indicate a genuine sense of curiosity about what reviewing would offer, and how they could improve their work by reading the reviews.

However, there was considerable anxiety about having to "show" their work to others and in the surveys anonymity was frequently mentioned as being important to their comfort:

> *I'm very glad my essay was anonymous. Totally hate the idea of anyone other than the lecturer/tutor reading my work and was considering losing the mark and not uploading my essay at all. Now that it's over though I guess it's good to get some feedback on ur work and to see how other people are tackling it.*

In the interviews, however, students were somewhat more open to the idea of not doing it anonymously.

There were few indications of students feeling intimidated by the technical requirements. A few software problems were encountered with the 2005 activity. Some students submitted documents in formats that were not supported by reviewers' computer systems (e.g., TIFF scanned images of hand-written notes), and there was a brief network outage. No problems were reported in 2006. Despite coming from non-technical backgrounds, the students were comfortable with using computers and with the Aropä interface.

The students were very positive about the marking rubric and its relation to the purpose of the assignment. They found the questions helpful, although a number did indicate that they would have preferred to have one open space where they could have commented generally. Having to re-state their assessee's argument and then articulate a counterargument seemed to be a valuable learning exercise. One student said that in re-stating the argument:

> *I didn't just copy things out from what they'd written... but tried to read and put it into my own words so that they would have an idea of what the reader was getting out of it. That was quite important. Yeah, I felt like I read them very carefully.being able to summarise the main points, putting it in your own words, being able to identify what the writer is trying to do... seeing the argument as a piece of writing that's been written in*

> *a certain way for a reason and then try to think about what that reason might be.*

She also stated that the counter-argument approach: "forces you to show that you've read it really carefully." She saw value in being able to separate out her own views on the argument by having to articulate the counter-argument, and had noted that point in class:

> *Yeah that's what having the counter-argument makes you do... even though I didn't believe in the counter argument, but I'd noted that down that we still had to do that, so I think that the exercise kind of forced you, its kind of important to do that.*

Students reported concerns about causing offence in the review but one responded that:

> *I think I was a lot nicer than the people who reviewed me! Instead of saying I don't care about your topic I said 'This would be interesting to people who... ' or 'I would be interested if I... ' Constructive criticism is good though.*

They largely expected to be good at reviewing, and were able to identify the skills needed for reviewing as: being critical, detached, analytical and logical. There was much interest in comparing their own performance to the rest of the class: "First essay reviewed made me feel good" (i.e., reassured); "I was relieved that others were not way better than mine."

Students were not involved in the marking but we discussed the issue in the interviews:

> JH: *You say that they were all really good. Would you be able to rank them?*
>
> Student: *Definitely... they were all pretty good. It'd be quite hard but if I went back to them maybe ... if I look back at them now at this stage I'd give them all As.*
>
> CK: *How did that make you feel, about your own essay?*
>
> Student: *It made me want to look at it more carefully... ( ) about the argument I was making and see how much research other people had done. I thought well, maybe I should go back in, like I had done research but I hadn't included it in the bibliography and everyone else had'it looked like maybe I'd hadn't read anything and made all this up, so it made me reassess what I'd done.*

There were very positive responses in having multiple perspectives on their writing. Their lecturer had indicated the importance of this as well. While being clear that he would be allocating the mark at the end, he said that he was:

> *...interested in the drafting mode... this way I'll save myself from looking at the drafts. I can't read 100 drafts. The idea of this is that you get a full response from other students. If it's just one to one, you'll get your usual range of A to D students, but this thing offers three to five responses for one paper. Across the five responses things will come up. This is a course on rhetoric in public culture. The groups of five act as the public audience.*

The students had completely 'taken on' this perspective. Comments on the value of the exercise were:

> *"(It's) to distribute a work to a wider public audience to test your rhetoric and better understand the audience you are trying to persuade. If you just write for a lecturer there is no feeling of audience."*
>
> *"The task of looking at one's own writing from an objective perspective in order to improve it or see things that have been missed is a hard thing to do. Sometimes we tend to get into a pattern of blindspots and it is therefore possible to read something a hundred times and still miss the same basic mistake every time. The value of both getting more than one set of other eyes on the writing as well as the experience of analysing and critiquing someone else's essay gives a much better insight in these aspects of own writing and writing in general."*

The marks for taking part in the exercise motivated many of the students we asked, but they also mentioned: mutual benefit; receiving feedback; skill training; improving their essay; feeling obligated to provide a review for other student who had done likewise; and curiosity.

In the interviews we asked students whether doing the reviewing had given them new insights into the marking that their tutors or lecturers did:

> CK: *Has it given you any more insight into the way your lectures or tutors might mark or approach the marking? Can you try and tell us what that might be?*
>
> Student: *Well, I guess you have to almost put yourself into their shoes, so you have to think if I was going to grade this, these are the things they would look at, I guess maybe look at the argument as a whole and kind of... I guess when writing it you feel you're really IN there and you're paying attention to every small detail and even though it might make sense in your own head when you read other people's and then look at theirs and then come back to yours, you're able to look at it with a slightly broader perspective.*

In response to the feedback they received, one participant reported that the consensus view of a post-assessment tutorial group was "better to be nailed by peers than by the lecturer." Another participant said "tutors can be 'beyond' where you're at; it's good to have reviewers at your own level." There was also an acknowledgement that lecturers don't always have time to give extensive feedback, and rarely look at draft work. Our study participants largely felt confident in rejecting misguided feedback, and in some cases adopted writing style or ideas from their peers.

There were no comments on the timing or the quantity of work required. In general we can say with confidence that the students reported being happy with the peer assessment process. Complaints largely related to not receiving all three reviews, due to other students in the class failing to fully participate in the exercise. The exercise was repeated in a largely identical manner in 2006. Although we did not interview students again, we did observe a significant increase in written comments. This coincided with the introduction of a WYSIWYG Javascript editor into Aropä that made it easy for reviewers to add section headings, font styles, itemised lists, smiley faces, etc. to their comments.

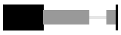
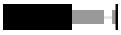
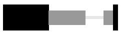
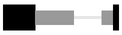
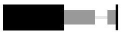
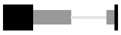
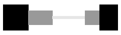
The web interface was easy to use	
The system responded quickly, and did not leave me waiting	
I feel comfortable with other students reading my work	
I feel comfortable assigning marks for other students	
Marking other students' work helps me spot mistakes in my work	
The comments from other students were helpful to me	
I would like more assignments to be peer marked	

Table 3: Likert scale survey questions and results for Computer Science. The graphics show the distribution of responses, Agree on the left and Disagree on the right, darker for Strongly.

4.3 Computer Science

Aropä has been extensively used in our introductory programming course for three years. The course enrols around 400 students in the first and second semesters, with a further 200 students taking the course in the summer semester. Peer assessment has been made a routine component of the courses. After each programming assignment, students are expected to complete three or four reviews, one of which is now typically a sample solution. The grading rubric is the same as used by the course markers, who also use the Aropä system. Marks are awarded for participating in the peer review, with some allowance made for the quality of their reviewing. The grade for the programming task is determined solely by the course markers.

An anonymous survey asked participants to respond with (strongly) agree, neutral, (strongly) disagree to a set of questions (see table 3), and for their open-ended comments on:

- What did you like most about the peer marking system?
- In what ways could the peer marking system be improved?
- Any other comments you would like to make.

The survey attracted just under 300 responses from 155 individuals. Most of the suggestions for improvement concerned interface and functionality issues that have since been addressed. Many students commented that the workload imposed by the peer assessment activity was too high, often taking longer to complete than the assignment.

> *1 or 2 assignments to mark 5 is waaay to much time. (i spent 1.5–2 hrs)*

Others felt they learned all they were going to from writing the first few reviews:

> *we need to mark 6 other student's assignments is too much, waste lots of time and learning nothing, I think marking 2 assignments would be enough*

Several issues were identified with the grading rubric:

> *More choices needs to be available, the "grades" that I can choose from does not include all possible errors that can be made.*
>
> *comments should not be necessary when all the assessment is correct*

mentions	issue
103	interface issues
47	workload is too high
28	problems with the grading rubric
18	dubious competence of peers
5	provide sample solutions
4	discomfort in writing comments
4	non-specific dislike
1	lack of feedback
1	inappropriate activity

Table 4: Summary of the open-ended improvement feedback from the Computer Science survey

> *Parts of the marking schedule were quite brutal... for example... int Q2 one might perform 1 wrong spelling but the rest is perfect, yet receive no marks*

There were some concerns with the competence of their peers:

> *Comments i recieved i found to be largly irrelevant and some left me questioning my peers' ability to accuratly mark my work.*

The remainder included suggestions for incorporating sample solutions in the review allocations, feelings of discomfort in reviewing, complaints about the lack of feedback received, and some non-specific indications of dislike. One student commented they felt the activity was inappropriate. The results are summarised in table 4.

The survey participants suggested a number ways in which they felt the peer assessment activity was of benefit. The most commonly mentioned was simply being exposed to a variety of coding styles:

> *You get a whole lot of different viewpoints to the assignment solution and it makes you think of all the other possibilities you could have used.*

Learning occured both from exposure to good code and exposure to mistakes:

> *So that I can see the mistakes that people make so I can improve myself. Also by looking at the "tops" coding can help me learn.*
>
> *Felt like a waste of time in the beginning, but it really is an excellent way of improving. I didn't realise the value of good comments and easy-to-understand code till I got a bunch of hard-to-understand code to mark. It really is very helpful.*

The material learnt was both specific to the assignment and more general:

> *I learned very quickly the mistakes that I had made in my own programs, when marking the other people's work*
>
> *It aided in the learning of certain aspects such as style and code conventions.*

The feedback from reviewers was variable, with some students writing detailed comments and others providing only shallow or empty responses. The students allocated conscientious reviewers were generally grateful for the feedback:

> *you get feedback from fellow students, which is interesting to read*

Comparing their own performance to that of their peers was another frequently mentioned benefit. This ranged from plain curiosity, to confidence building, through to awareness of an audience (a theme noted much more strongly in the English class):

mentions	like
40	exposure to a variety of coding styles
32	learning examples of good coding
32	non-specific positive comment
35	helpful feedback received
24	the system was convenient and easy to use
20	learning to identify poor programming constructs and mistakes
19	improving (debugging) their own code for this assignment
19	comparing own performance to peers
15	helpful reading code
10	helping others by giving feedback
5	learning by marking
2	the exercise motivated them to work harder
2	gaining an insight into the marking process
2	anonymity relieved concerns about fairness

Table 5: Summary of responses to "What did you like most about the peer marking system?"

> *Discovering the level of understanding of program writing of other students*
>
> *The oppertunity to guage my work against that of other sutdents.*
>
> *. . . got me thinking more about my code that I had previously — especially towards readability and reasonably named variables. Suddenly the use of 'foo' as a variable name isn't as effective when several people are looking at your code!*

The results are summarised in table 5.

One (unexpected) consequences of the routine use of peer assessment in this course was the ability to identify capable student reviewers, who were then offered employment as course markers in the following semester. The lecturers observed that complaints about marking reduced very significantly, from between twenty and forty after each assignment to virtually none. This was felt to be due to students' increased awareness of the marking process, and more capable markers being employed. The use of Aropä by the course markers also allowed their progress to be monitored by the lecturer and tardy markers chased up.

5 Summary and conclusions

Hanrahan & Isaacs (2001) have noted that although the range of general studies on peer assessment is growing and some general trends can be noted, peer assessment has to be implemented on a case-by-case basis in varying subjects and contexts. They have commented that "case-based literature in this area is still alarmingly sparse." The development of Aropä and our study of its uses in three different courses at the University of Auckland, using the same web-based tool, enable us to make some comparisons and draw some conclusions.

Aropä has been deployed in a very wide range of contexts, and has proven acceptably versatile. Students reported that peer assessment contributed to learning at many different levels, as a consequence of:

- perceiving writing as a public activity;
- reflecting on the grading rubric, which invites questions of "what is important;"
- exercising judgement in awarding marks;
- encountering examples of good style or technique;
- encountering poor solutions that reveal mistakes to avoid;
- altruistic feelings of serving a valuable community role;
- receiving accurate, constructive, timely feedback;
- selecting between worthwhile and misguided feedback.

We observed very different reactions from students in the three subjects. Pharmacology students showed the most resistance. Computer Science by-and-large accepted the activity. In English, it was positively received as a natural part of the course. The reasons for this are complex and will require further analysis, but some discussion is provided in what follows.

Boud et al. (1999) point out that "assessment is the single most powerful influence on learning in formal courses, and if not designed well, can easily undermine the positive features of an important strategy in the repertoire of teaching and learning approaches." In summarising the literature in the field of assessment they draw up the following effects of assessment on learning.

- Individuals are emphasised

They argue that the overriding paradigm of assessment is that it is an "individual and competitive" process in most institutions. While there has been something of a shift to criterion-referencing, norm-referencing still dominates. If individualistic views of assessment are dominant, collaboration can be seen as 'cheating.'

- Assessment exercises power and control over students

They state that assessment is the principal mechanism whereby staff exercise power and control over students. The effect of this on learning is to circumscribe it to the range of outcomes unilaterally defined as legitimate by staff. Students learn first to distrust their own judgements and then act as agents to constrain themselves.

- Assessment exerts a backwash effect on learning.

The authors cite the work of Marton et al. (1997) and they argue that "inappropriate forms of assessment appear to encourage students to take a surface approach to learning. . . conforming to the narrowest interpretations of assessment tasks and working to beat the system rather than engage in meaningful learning."

- Overload of tasks discourages deep approaches to learning.

Drawing on Ramsden & Entwistle (1981), Boud et al. (1999) argue that overloading contributes to students taking a surface approach to learning tasks, and caution that overloading can lead to peer learning tasks being either ignored or falling into disrepute.

- Assessment practices need to be matched to outcomes.

They argue that outcomes need to be designed in terms of basic knowledge, understanding, communicative and competency aims being pursued in a course.

- Formal assessment processes should encourage self-assessment

They point out that assessment in higher education has a dual function of judging for the purpose of providing credentials and for the purpose of improving learning. With regard to the latter, they claim that assessment should leave students better equipped to engage in their own self-assessment.

The authors also argue that there are both pragmatic and principled reasons for the current focus on peer learning in university courses. Yorke (2003), writing about formative assessment, identifies these pressures as including:

- An increasing concern with attainment standards, leading to greater emphasis on the (summative) assessment of outcomes;
- Increasing student/staff ratios, leading to a decrease in the attention given to individuals;
- Curricular structures changing in the direction of greater unitisation, resulting in more frequent assessment of outcomes and less opportunity for formative feedback;
- The demands placed on academic staff in addition to teaching.

At the University of Auckland these pressures take the form of pressures to both intensify teaching work and at the same time improve quality, in relation to criteria like the fostering of critical and independent thought, creativity and imagination, communication and research skills.

It is clear that the students' uptake of peer assessment in our study reflects these pressures and trends in complex and contradictory ways. The design of the peer assessment tasks across the three cases provides insights into the role of peer assessment in relation to these general trends.

The Pharmacology students (who were required to provide marks for each other worth 20% of the total mark) responded (in a more aggrieved way) to what they believed were the pragmatic reasons; i.e., suggesting that they were having to do the lecturers' job, but realising the benefits for learning as they engaged with the task. The responses also suggested that they felt overloaded at the time of doing the task and became preoccupied with concerns over the fairness of the system.

Much less was at stake with the Computer Science and English assignments. Computer Science students could see some personal benefits in using the system. In English, the students reflected an awareness of the public role of their work and recognised the pedagogical value of the close matching of the assessment with the outcomes and realised the demands of the task as they internalised this matching. It must be noted that their task was formative, implicitly led to self-assessment and only counted for an ungraded 5%.

It may be, therefore, that the reversal of power relations implied in the case of the Pharmacology students was a little too much, too soon and that it is important to recognise that the dominant assessment paradigm to some extent, constructs students' responses. A lecturer in Computer Science put this point well:

> *There's a difficulty primarily in marketing it. Getting students on board is difficult. It has to be done carefully, it needs to be approached not as an evaluative tool but pitched as a way of learning. The students see the potential as long as its pitched appropriately.*

The issue of the assessment exerting a backwash effect links closely with the matching of the assessment tasks and the outcomes which we suggest were seen by the lecturers in our study as dilemmas. The English lecturer's comment that writing is something students learn across the years of their time at university, not in a one-off course or task, is important. It is possible that the definition of outcomes for courses and for assessments can give out triumphalist messages about what one exercise can achieve, while at the same time the construction of, and the language used in rubrics potentially promotes inappropriate or reductionist ideas about the matching of assessment and outcomes.

Our data seems to suggest that students are very aware of these issues. We find it heartening to see students' ability to read in between the lines of these pressures on their lecturers. They respond both on the pragmatic terrain, but also indicate their desire for deep learning and reflection. The very sophisticated comment made by a Pharmacology student speaks volumes about this backwash effect and the possible constraints of the rubrics in giving expression to the matching of the assessment tasks and the outcomes:

> *My assignment would be quick and simple to mark, ordered according to the marking schedule, rather than the assignment outline, and I could make sure that everything the marking schedule required was in my assignment, and nothing more. No extra research, nothing interesting; i.e., use bullet point format to create a piece of work designed only to meet the requirements of a marking schedule.*

A Computer Science lecturer specified the problems with the rubrics clearly:

> *I'd recommend to others now to try and make rubrics that are reasonably vague and general. When you are too specific you eliminate some of the critical thinking and it becomes a mechanical process. Some of the early rubrics we built, we could have got a computer programme to mark them! In writing code, the style is much greater than the sum of the parts. Sometimes the code meets the specification, but it's just being done badly — it's very difficult to write a rubric to capture that. If it's very open, then it can help if the students are able to discuss the marking criteria, and we use a wiki so that they can do this.*

The English lecturer's decision to engage his students with both humorous and reflective questions provides a sophisticated solution to these dilemmas, and the students responded very well to it.

Finally, we do not have clear-cut data to back up this claim, we gained the impression that the organization of the peer assessment exercises through the automated tool of Aropä lent a certain formality and anonymity to the tasks, which gave peer assessment greater credibility. A number of students had commented on how they had done peer assessment informally in tutorials by exchanging their written work with fellow students. Many limitations to, and problems with this were indicated. Aropä 'puts' their work into a new and neutral context, and we suggest that this in itself is helpful if peer assessment is to be taken up more widely.

6 Acknowledgements

We would like to thank our interview subjects for sharing their views so openly. The anonymous reviewers provided extensive, thoughtful comments, for which we are grateful. Particular thanks to Paul Denny, Jeff Keelan, Andrew Luxton-Reilly and Stephen Turner for agreeing to use Aropä in their courses.

References

Ballantyne, R., Hughes, K. & Mylonas, A. (2002), 'Developing procedures for implementing peer assessment in large classes using an action research process', *Assessment & Evaluation in Higher Education* **27**(5), 427–441.

Barrie, J. (2006), '`http://www.turnitin.com/`'. Accessed 10 August 2006.

Bloom, B. S., ed. (1956), *Taxonomy of educational objectives: The classification of educational goals. Handbook I, cognitive domain*, Longmans, Green, New York; Toronto.

Boud, D., Cohen, R. & Sampson, J. (1999), 'Peer learning and assessment', *Assessment and Evaluation in Higher Education* **24**(4), 413–426.

Dochy, F., Segers, M. & Sluijsmans, D. (1999), 'The use of self-, peer and co-assessment in higher education: A review', *Studies in Higher Education* **24**(3), 331–350.

Hamer, J., Ma, K. T. & Kwong, H. H. (2005), A method of automatic grade calibration in peer assessment, *in* A. Young & D. Tolhurst, eds, 'ACE'05 Australasian Computer Society Education Conference', Vol. 42 of *Conferences in Research and Practice in Information Technology*, Australian Computer Society, pp. 67–72.

Hanrahan, S. & Isaacs, G. (2001), 'Assessing self- and peer-assessment: The students' views', *Higher Education Research and Development* **20**(3), 54–66.

Marton, F., Hounsell, D. & Entwistle, N., eds (1997), *The Experience of Learning: Implications for Teaching and Studying in Higher Education*, 2nd edn, Edinburgh: Scottish Academic Press.

Moxiecode Systems AB (2006), '`http://tinymce.moxiecode.com/`'. Accessed 10 August 2006.

Ramsden, P. & Entwistle, N. (1981), 'Effects of academic departments on students' approaches to studying', *British Journal of Educational Psychology* **51**, 368–383.

Topping, K. (1998), 'Peer assessment between students in colleges and universities', *Review of Education Research* **68**(3), 249–276.

Yorke, M. (2003), 'Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice', *Higher Education* **45**, 477–501.